Д. В. Сичинава

НАЦИОНАЛЬНЫЙ КОРПУС РУССКОГО ЯЗЫКА: ОЧЕРК ПРЕДЫСТОРИИ

Странной может казаться сама идея написать историю предприятия, которое существует «в публичном формате» лишь около полутора лет. Действительно, лишь 29 апреля 2004 года Корпус стал доступен на Интернет-сайте для всех желающих, а не только для посвящённых, кому были знакомы капризно меняющиеся время от времени «логин» и «пароль». И история его — это, в определённой степени, и есть его современность. Но открытию предшествовали три с лишним года достаточно интенсивной подготовки. Это был очень полезный практический опыт корпусного строительства, тем более ценный, что осуществлялся не по раз заданному плану и теории, которую надлежало только воплотить, а был неотделим от постоянного поиска — вплоть до того, что, с одной стороны, от твёрдо однажды принятых решений приходилось отказываться, а с другой стороны, чуть ли не каждую неделю одно время возникали новые, всё более заманчивые направления деятельности, разумеется, в рамках Корпуса. И этот опыт, как мне кажется, небезынтересен не только участникам проекта.

В начале работа по подготовке корпуса была, так сказать, невидима — ее планы обсуждались на различного рода московских лингвистических семинарах (с участием далеко не только москвичей), но период открытого доступа к результатам этой работы: доклады на конференциях, появление работ с ссылками на Корпус, а главное, фиксируемые на беспристрастных счётчиках сотни посещений сайта в день, письма пользователей из Воронежа и Томска, Нью-Мексико и Прованса, Египта и Австралии, — был ещё впереди. Об этом предварительном этапе мне бы и хотелось подробно рассказать.

Всё началось с Центра лингвистической документации. Так назывался кружок московских лингвистов под руководством В. А. Плунгяна и М. А. Даниэля, проводивший цикл семинаров в Центре непрерывного математического образования по приглашению руководителей центра. Инициатива, как и во многих других лин-

гвистических начинаниях, исходила от математиков — И. С. Красильщика и С. К. Ландо. Идея создания кружка состояла в том, чтобы объединить разрозненные усилия лингвистов и программистов по созданию исследовательских инструментов нового поколения: баз данных, программ автоматического поиска, интеллектуальных систем, а также — корпусов текстов.

Первый семинар состоялся 13 октября 2000 года, и на нём уже обсуждались основные темы, которые стали лейтмотивом дальнейшей деятельности кружка. Говорилось о том, что несмотря на давнее и постоянное внимание отечественных лингвистов к проблеме компьютеризации и информатизации своей науки — прежде всего, к удобному хранению языковых данных в компьютерном виде — в этой области имеется до обидного мало результатов, доступных всему лингвистическому сообществу. Из средств работы с лингвистическим материалом основным, причём достаточно давним, но неустаревающим достижением стала СУБД Starling (о которой на семинаре ЦЛД 9 февраля 2001 г. делал доклад безвременно покинувший нас С. А. Старостин). Существовал и ряд других доступных лингвистам средств хранения и поморфемной нотации (глоссирования) текстов и словарей, программ хранения аннотированной звучащей речи и т. п. Всё это обсуждалось на семинарах, но и достоинства, и недостатки названных программных средств отступали на второй план перед следующим капитальным фактом. В отечественной прикладной лингвистике зияла парадоксальная брешь — по сути, не было общедоступного представительного и аннотированного (размеченного) *корпуса* русского языка, с которым можно было работать лингвисту.

Отечественная корпусная лингвистика началась как минимум с 1980-х годов, с работ А. П. Ершова и В. М. Андрющенко, посвящённых тому, «каким должен быть» машинный фонд русского языка. Тогда же в Институте русского языка имени В. В. Виноградова РАН был создан Отдел машинного фонда, занимавшийся, в частности, переводом русских литературных текстов в электронную форму. К сожалению, технических и организационных возможностей для того, чтобы создать полноценный машинный фонд текстов русского языка, в тот период не хватило. (Надо сказать, что, тем не менее, многие материалы, собранные и обработанные в Отделе машинного фонда, особенно по авторам XIX века, впоследствии удалось использовать и при работе над нашим корпу-

сом, так что вклад сотрудников этого отдела в Национальный корпус русского языка также существен.)

Поскольку единого авторитетного для всех корпуса русских текстов создать тогда не удалось, едва ли не у каждого лингвиста, серьёзно интересовавшегося корпусными исследованиями, имелась своя, отличная от прочих, электронная коллекция русских текстов, которой он, может быть, и был готов поделиться со всеми желающими. Но проблема состояла в том, что в отсутствие какого-либо организационного центра и «правил игры» такое «локальное» распространение «кустарных» продуктов (зачастую небольших по объёму) было малоэффективно.

Кроме того, тексты, почерпнутые из открытых источников, изобиловали, как нередко бывает, искажениями разного рода; откуда они взялись, кто за них отвечает, как сделать так, чтоб эта ошибка прекратила кочевать из коллекции в коллекцию — в такой ситуации было нельзя выяснить в принципе. Помню случай, когда, воспользовавшись электронным текстом из такой коллекции, один замечательный лингвист посвятил несколько абзацев своей работы анализу нестандартного употребления слова *вина* у Булгакова, где фигурировал *не чуявший вины ресторан*. К счастью, уже на стадии корректуры бдительный редактор сам *почуял* неладное, обратился к печатному тексту Булгакова и выяснил, что никакой *вины* там просто нет — ресторан не чует всего-навсего *беды*; возникший из-за причуд программы распознавания (но, конечно же, никак не по *вине* автора) пассаж успели вычеркнуть. Общедоступные корпуса русского языка, созданные зарубежными славистами, конечно же, были (Уппсальский и Тюбингенский, к тому времени уже объединившиеся), но они сильно ограничены по объёму, отчасти устарели по представленным текстам, и, опять-таки, не содержали морфологической разметки.

Уже на втором семинаре ЦЛД в октябре 2000 г. С. А. Шаров — один из тех, кто впоследствии предложил саму идею создания представительного Национального корпуса русского языка — сделал доклад на тему «Введение в проблематику корпусной лингвистики и форматов хранения текстов (в первую очередь, русских литературных текстов)». Около нового 2001 года по инициативе В. А. Плунгяна началось уже и формирование коллектива, занимавшегося будущим русским корпусом. С самого начала ни у кого не вызывало сомнений, что этот корпус не станет частным достоя-

нием нашего центра, но будет размещён в Интернете и доступен всем желающим. Предполагалось, что основной (если не единственной) его разметкой станет морфологическая, а грамматическая омонимия при этом будет снята. Кроме того, корпус тогда ещё не мыслился как репрезентативный по отношению к генеральной совокупности русских текстов: в него должны были войти прежде всего литературные тексты, написанные без нарочитого «экспериментаторства» и культурно значимые. Помню, долгое время шло обсуждение в основном списка авторов, носившее (на первый взгляд) неожиданно нестрогий филологический характер: можно ли считать «хорошим русским языком», допустим, язык Гроссмана или Астафьева (тогда в «пилотный» корпус не вошли ни тот, ни другой, но уже на следующем этапе, о котором ниже, оба писателя в Национальном корпусе появятся). У В. А. Плунгяна, М. А. Даниэля, А. А. Егорушкина, меня и некоторых других (обсуждавших эту идею в первые месяцы 2001 г.) возникали варианты названия вроде «Стандартный корпус русского языка», «Представительный корпус русского языка» (под «представительностью» мы тогда имели в виду довольно субъективную «нормативность»), «Русский стандарт» (в чём была изрядная доля шутки — так называлась активно рекламировавшаяся в то время водка, а на Воздвиженке около Института языкознания висел рекламный щит этого напитка, который нас во многом вдохновлял: «РУССКИЙ СТАНДАРТ: СДЕЛАТЬ НЕВОЗМОЖНОЕ»). В конце концов рабочим названием стало «Корпус ЦЛД—МГУ». Действительно, в судьбе корпуса принял участие не только с административной точки зрения слабооформленный (хотя организационно активнее иных академических заведений) ЦЛД, но и кафедра теоретической и прикладной лингвистики МГУ им. М. В. Ломоносова; в конце января 2001 г. на ней было принято решение способствовать созданию корпуса и привлекать студентов отделения ТиПЛ и аспирантов кафедры к работе над ним.

С другой стороны, технически-программистскую поддержку корпус получил со стороны компании «Яндекс», ранее уже участвовавшей в лингвистических проектах и заинтересованной в корпусе как в важном средстве совершенствования целого ряда программ автоматической обработки текста; тогда же состоялись первые встречи участников ЦЛД с техническим директором компании И. В. Сегаловичем и программистом В. А. Титовым, которые до сих пор являются ключевыми участниками работ по Корпусу. В марте

2001 года к коллективу присоединился А. Е. Поляков (кстати, познакомились мы с ним на семинаре ЦЛД), — единственный на тот момент в России лингвист, имевший опыт создания свободно доступного объёмного ресурса с полностью однозначной морфологической разметкой и возможностью поиска по ней — это «Словарь языка Грибоедова» (http://www.inforeg.ru/electron/concord/concord.htm). Сотрудничество облегчало то, что А. Е. Поляков и ранее пользовался средствами морфологической разметки, созданными «Яндексом» при участии сотрудников Лаборатории компьютерной лингвистики ИППИ РАН (программа Mystem); им также был разработан шаблон для правки разметки в текстовом редакторе и интерфейс поиска. И в начале лета уже начались работы по снятию морфологической омонимии в небольшом — тогда ещё даже менее 1 млн. словоупотреблений — корпусе, состоявшем из художественных текстов второй половины XX в. (всего несколько авторов: Аксёнов, И. Грекова, Трифонов, Довлатов, Татьяна Толстая). Снятие омонимии (тогда оно было полностью ручным) финансировал «Яндекс», а работали студенты и аспиранты МГУ и РГГУ.

В сентябре 2001 г. мы с И. В. Сегаловичем и А. Е. Поляковым сделали доклад о корпусе на очередном после летних каникул семинаре ЦЛД — это был уже не широковещательный прожект, а отчёт о действительно проделанной работе. Данный этап деятельности был отражён в моей статье, которую тогда же и отдали в печать, но к моменту выхода она, что достаточно показательно для истории нашего Корпуса, успела сильно устареть[1]. В довольно кратком тексте я описал и принцип отбора текстов, и морфологическую концепцию, и технологическую цепочку, и устройство разметки; как сказал по схожему поводу один из персонажей «Мемуарных виньеток» А. К. Жолковского — «не может в одном и том же докладе быть и про лампы в ЭВМ, и про порядок слов». Но для того этапа это было еще естественно — корпус был небольшой, «пилотный», и слишком мало задач пока что ставилось.

Качественно новый этап начался с 2002 года, когда расширилась и концепция корпуса, и его программная база, и организационные «площадки». Сначала несколько слов о последнем факторе,

[1] Д. В. Сичинава. К проблеме создания корпусов русского языка // Научно-техническая информация, серия 2, № 11, 2002, с. 25-31; несколько сокращенный вариант, с бóльшим упором на Интернет-специфику: Русский язык в Интернете. Сборник статей. Казань, «Отечество», 2003, с. 111-122.

может быть, не самом интересном, но без которого не может быть полноценной работы. В ВИНИТИ на базе группы Е. В. Падучевой был организован Отдел лингвистических исследований (руководимый в настоящее время Е. В. Рахилиной), занявшийся корпусом как одной из плановых тем в тесном сотрудничестве с Институтом русского языка им. В. В. Виноградова РАН — там к работам подключился уже упоминавшийся Отдел машинного фонда. Так Корпус «институционализировался» в обоих названных учреждениях. В частности, это означало постоянно действующие семинары с заранее объявленной повесткой дня и протоколом обсуждения, которые проводятся попеременно в обоих институтах, длятся по несколько часов подряд и собирают от 10 до 40 человек участников.

Вообще говоря, имеется известный тезис, согласно которому коллективное предприятие, обзаведшееся наконец «кабинетом и табличкой», обречено на деградацию. Но, к счастью, специфических «кабинета и таблички» у Национального корпуса нет до сих пор — и, по-видимому, не может быть: и в самом начале, и теперь Корпус придумывается и делается руками лингвистов самой разной ведомственной принадлежности — при неизменной моральной, технической и административной поддержке директора Института русского языка А. М. Молдована.

Теперь о материальной стороне дела: бюджете. Корпус получил поддержку по ряду грантов. Первым, как я уже сказал, поддержку нашему начинанию оказал «Яндекс»: деньги были потрачены на разметку «пилотного» корпуса с помощью программы И. В. Сегаловича и на снятие в нем омонимии. Потом наступил финансовый перерыв — в это время мы решили попробовать перейти на другую программу морфологического анализа, которая базировалась на грамматическом словаре А. А. Зализняка, — Dialing, создававшейся А. В. Сокирко и Д. В. Панкратовым (изначально — для системы, которая делалась под руководством Н. Н. Леонтьевой). Но всякую, даже очень хорошую программу, надо «дотягивать» и приспосабливать — и мы гордимся, что, совершенствуя снятие омонимии, так сказать, в новых условиях, мы работали бок о бок с этими программистами. Тут нам помог молодежный грант Президиума РАН (ведь омонимию снимали в основном студенты и аспиранты!).

Интересы создателей Корпуса быстро вышли за пределы компактного корпуса текстов «хороших писателей». В этом немалое содействие нам оказал С. А. Шаров. Он предложил трансформа-

цию нашего проекта в «Национальный корпус русского языка» (тогда так не называвшийся) — аналог европейских «национальных корпусов» (таких, как Британский или Чешский), представляющих данный язык во всем его разнообразии на определённом историческом отрезке. Такая задача диктовала и репрезентативность корпуса, и его большой объём (порядка 100 млн. словоупотреблений), и невозможность ограничиваться ручным снятием омонимии. Эта задача, как уже говорилось, была поддержана программой Академии наук «Филология и информатика». Летом 2002 г. была разработана концепция сосуществования двух основных подкорпусов — с полностью снятой грамматической неоднозначностью (меньшая часть) и со всеми теоретически возможными разборами (большая часть).

Программа РАН «Филология и информатика», в рамках которой и был осуществлен Национальный корпус, открылась зимой 2003 года. Программа началась с заседаний, на которые мы пришли уже как специалисты с практическим опытом корпусной лингвистики, имеющие значительный задел (полумиллионный корпус с морфологической разметкой) и мечту — создать большой представительный корпус русского языка.

Удачным образом, в необходимости осуществления нашего замысла долго убеждать никого не пришлось. Все участники этих обсуждений — Ю. Д. Апресян, Н. Н. Казанский, В. Б. Касевич, А. М. Молдован, А. Я. Шайкевич и др. — не сомневались в том, что такой инструмент для изучения русского языка будет крайне полезен. В. Б. Касевич, много занимавшийся проблемами корпусной лингвистики, предложил закрепить в названии проекта термин «Национальный корпус русского языка», по принятым в мировой практике образцам. Проект под таким названием был одновременно подан петербургскими коллегами под руководством Л. А. Вербицкой (и с нашим участием) и в Российский гуманитарный научный фонд. Поддержка этого фонда позволила нам подготовить к размещению в Интернете и снять омонимию с небольшого массива разговорных текстов.

Концепция Корпуса расширялась в разных направлениях. Первоначальный проект морфологической разметки был достаточно «эзотерическим» с точки зрения теоретических принципов (как и во многих других аннотированных корпусах). Предполагалось, например, что разграничение по признаку «часть речи» будет прово-

диться лишь в наиболее очевидных и морфологически бесспорных случаях: например, для большого класса неизменяемых слов, для которых словари дают омонимию (или, если угодно, синтаксическую полифункциональность) вида «частица/союз», «частица/наречие» и т. п. (*же, словно, ещё* и др.) наш корпус давал бы во всех случаях единый разбор через дробь. Более того, не предрешался и вопрос о статусе словоформ на *-о/-е* типа *весело* (наречия, краткие прилагательные или предикативы) или субстантивированных прилагательных вроде *военный* или *дежурная*. Предполагалось, что с точки зрения морфологии такие различия не существенны, и пользователь сам сумеет провести в своих целях релевантные границы. В соответствии с этими принципами была начата разметка вышеуказанной «пилотной» части корпуса.

Но со временем обнаружилось, что такой подход не встречает широкого понимания: большинство пользователей — как соотечественники, привыкшие к «школьной» грамматической номенклатуре, так и изучающие русский язык слависты — нуждаются в более чётких ориентирах, фиксирующих определённую традицию. Вот почему мы всё-таки стали работать с традиционной классификацией частей речи (в версии, отраженной в «Грамматическом словаре» А. А. Зализняка, включая даже такие теоретически небесспорные классы, как «вводные слова»). Кое-что при этом пришлось поправить автоматически или даже переразметить.

Таким образом, постепенно формировался нынешний «морфологический стандарт» корпуса. Различные варианты этого стандарта неоднократно обсуждались с коллегами-экспертами из Лаборатории компьютерной лингвистики ИППИ РАН и Санкт-Петербургского университета. В частности, по предложению группы В. Б. Касевича (по итогам обсуждения морфологического стандарта с ним и его сотрудниками А. В. Венцовым и Е. В. Ягуновой) в подкорпус со снятой омонимией была введена акцентная разметка и буква «ё», а также отсутствующая в «Грамматическом словаре» Зализняка часть речи «местоименное наречие».

С другой стороны, по инициативе и при активном участии С. А. Шарова была разработана подробная система классификации текстов (по-видимому, одна из самых детальных в мировой практике) — «метаразметка». Речь идет об особом типе информации, приписываемой тексту в целом, при помощи которой пользователь может искать тексты, созданные в определенный период, на опре-

деленную тему, определенного жанра и т. п., легко по мере надобности формируя свой рабочий подкорпус. Разумеется, нельзя не упомянуть неоценимый вклад в разработку нашей системы метапризнаков, внесённый А. Е. Поляковым и С. О. Савчук, которая тоже начала работать с нами на этом этапе, перейдя на работу в Отдел машинного фонда русского языка и взяв на себя дополнительно важнейшие организационные функции по обеспечению проекта. Фактически, именно С. О. Савчук является главным координатором работ по Корпусу, она объединяет людей из разных институтов, городов и даже стран (ведь в создание корпуса внесли значительный вклад и российские специалисты, временно работающие за рубежом), регулируя потоки размеченных и неразмеченных текстов, официальной документации, командировочных удостоверений, финансовых отчетов, канцтоваров и т. п. Надо сказать, что не только такие активные участники проекта, как С. О. Савчук и Е. А. Гришина, но и все остальные сотрудники Отдела Машинного фонда русского языка под руководством А. Я. Шайкевича влились в работу над Корпусом. В частности, все тексты XIX века, собранные в Отделе (а это несколько миллионов словоупотреблений, включая полное собрание сочинений Ф. М. Достоевского, тексты А. А. Бестужева-Марлинского, В. В. Крестовского и др.), были переданы в ИЛИ РАН в Санкт-Петербург и вошли в фонд Национального корпуса: работу по формированию и разметке подкорпуса текстов XIX в. взяла на себя группа молодых сотрудников ИЛИ РАН под руководством Н. Л. Дич.

Вообще, по разным поводам не раз приходилось говорить, что Корпус — это прежде всего люди. Энтузиазм, идеи (разной степени «безумности») и просто человеческое участие и помощь. Число людей, имеющих отношение к этому проекту, непрерывно растет. Мы рады тому, что проект притягивает людей, рады появлению новых коллег — из Крыма, Кемерова, Перми, Саратова, Томска — и прочным связям со старыми, в числе которых хотелось бы назвать редактора издательства «Вагриус» Е. Д. Бычкову, сотрудника информационного агентства «Интегрум» Л. М. Гершензона, директора научно-методического центра по компьютерной лингвистике Воронежского университета проф. А. А. Кретова, доцента кафедры математической лингвистики Санкт-Петербургского университета И. В. Азарову, зав. отделением славянских и балтийских языков и литератур Хельсинкского университета, проф. А. Мустайоки и сотрудников этого отделения Е. Ю. Протасову и М. В. Копотева.

С конца 2002 г. корпус со снятой омонимией стал доступен для поиска — пока ещё под паролем по адресу www.ruscorpora.ru, на сервере компании «Яндекс». Интересно, что буквально с того времени, как «Яндекс» зарегистрировал это доменное имя, к нему было приковано внимание деятелей Интернета — моментально в службу поддержки стал приходить поток писем: «что это?», «что это за новый проект Яндекса?» И в дальнейшем Национальный корпус привлекал интерес программистов, компьютерщиков и всех, кто интересовался развитием Рунета. Уже в первые месяцы после официального открытия Корпус стал предметом рецензий в профессиональной компьютерной среде, для которой теоретическая лингвистика представляла лишь косвенный интерес, на таких сайтах, как «Вебинформ». Информация о сайте, несмотря на отсутствие открытого доступа, распространилась среди лингвистов, причём не только в России, достаточно быстро; так, уже весной 2003 года он регулярно использовался в преподавании русского языка в Карловом университете Праги (в чём принял некоторое участие и автор этих строк, посетивший тогда традиционную Школу Матезиуса и ставший свидетелем огромного интереса к нашему проекту как со стороны математических лингвистов, создателей Чешского национального корпуса, таких, как В. Петкевич или Я. Гайич, так и русистов — преподавателей Восточнославянской кафедры философского факультета).

После пополнения корпуса новыми текстами, открытия раздела со снятой омонимией, исправления ряда ошибок, составления руководства пользователя и — the last but not the least — создания дизайна, выполненного выпускницей ОТиПЛа МГУ А. С. Зыковой (в конце того же 2003 г. студия, которую она представляла, победила во внутреннем конкурсе, когда мы ещё не знали, что имеем дело с коллегой) — сайт www.ruscorpora.ru уже мог стать достоянием широкой публики.

Днём открытия сайта, как уже говорилось, стало 29 апреля 2004 года. Незадолго до этого Корпус был представлен на конференциях, организованных филологическим факультетом МГУ и ИМЛИ РАН. С этого времени предыстория проекта заканчивается и начинается его история — как и положено для динамически развивающегося проекта, практически совпадающая с его современным состоянием. О том, что в Корпусе есть сейчас и что будет в ближайшем будущем — другие статьи этого сборника.